On the estimation of the Cheeger constant

Ery Arias-Castro (*UC San Diego*)

with

Bruno Pelletier (*Université Rennes II*)
Pierre Pudlo (*Université Montpellier II*)

Graph partitioning applied to clustering points in space

The clustering problem

Given observations $X_1, \ldots, X_n$, partition the sample into k groups:

- ▷ dissimilar groups;
- ▷ similar observations within each group.

Graph partitioning applied to clustering points in space

The clustering problem

Given observations $X_1, \ldots, X_n$, partition the sample into k groups:

- ▷ dissimilar groups;
- ▷ similar observations within each group.

Numerous existing techniques:

- hierarchical classification;
- k-means algorithm;
- level set methods;
- graph-partitioning heuristics.

Examples of graphs

- $X_1, \ldots, X_n$ i.i.d. valued in some subset $M \subset \mathbb{R}^d$.

Examples of graphs

- $X_1, \ldots, X_n$ i.i.d. valued in some subset $M \subset \mathbb{R}^d$.
- Define a graph $G_n = (V_n, E_n)$ with vertices $V_n = \{1, \ldots, n\}$.
  The graph may have weights (weight matrix $\mathbf{W}$).

Examples of graphs

- $X_1, \ldots, X_n$ i.i.d. valued in some subset $M \subset \mathbb{R}^d$.
- Define a graph $G_n = (V_n, E_n)$ with vertices $V_n = \{1, \ldots, n\}$.
  The graph may have weights (weight matrix $\mathbf{W}$).
- The graph represents a rough skeleton of M.

Examples of graphs

- $X_1, \ldots, X_n$ i.i.d. valued in some subset $M \subset \mathbb{R}^d$.
- Define a graph $G_n = (V_n, E_n)$ with vertices $V_n = \{1, \ldots, n\}$. The graph may have weights (weight matrix $\mathbf{W}$).
- The graph represents a rough skeleton of M.

▷ **ε-ball graph**

$$i \sim j \quad \text{if} \quad \mathrm{dist}(X_i, X_j) \leq \varepsilon.$$

Examples of graphs

- $X_1, \ldots, X_n$ i.i.d. valued in some subset $M \subset \mathbb{R}^d$.
- Define a graph $G_n = (V_n, E_n)$ with vertices $V_n = \{1, \ldots, n\}$.
  The graph may have weights (weight matrix $\mathbf{W}$).
- The graph represents a rough skeleton of M.

▷ **ε-ball graph**

$$i \sim j \quad \text{if} \quad \mathrm{dist}(X_i, X_j) \leq \varepsilon.$$

▷ ***k*-nearest neighbor graph**

$$i \sim j \quad \text{if } X_j \text{ is one of the } k\text{-nearest neighbors of } X_i.$$

Examples of graphs

- $X_1, \ldots, X_n$ i.i.d. valued in some subset $M \subset \mathbb{R}^d$.
- Define a graph $G_n = (V_n, E_n)$ with vertices $V_n = \{1, \ldots, n\}$. The graph may have weights (weight matrix $\mathbf{W}$).
- The graph represents a rough skeleton of M.

▷ **ε-ball graph**

$$i \sim j \quad \text{if} \quad \text{dist}(X_i, X_j) \leq \varepsilon.$$

▷ ***k*-nearest neighbor graph**

$$i \sim j \quad \text{if } X_j \text{ is one of the } k\text{-nearest neighbors of } X_i.$$

▷ **Fully connected weighted graph**

$$\text{For example}: \; w_{ij} = \exp\left(-\text{dist}(X_i, X_j)^2 / h^2\right)$$

Normalized cut and Cheeger constant

Bipartite graph cut problem

Split the graph $G_n = (V_n, E_n)$ into S and S^c, with $S \subset V_n$.

Normalized cut and Cheeger constant

Bipartite graph cut problem

Split the graph $G_n = (V_n, E_n)$ into S and S^c, with $S \subset V_n$.

For S a subset of the graph, define

$$\sigma(S) = \sum_{i \in S} \sum_{j \in S^c} w_{ij} \quad \text{discrete perimeter}$$

$$\delta(S) = \sum_{i \in S} \sum_{j \neq i} w_{ij} \quad \text{discrete perimeter}$$

Normalized cut and Cheeger constant

Bipartite graph cut problem

Split the graph $G_n = (V_n, E_n)$ into S and S^c, with $S \subset V_n$.

For S a subset of the graph, define

$$\sigma(S) = \sum_{i \in S} \sum_{j \in S^c} w_{ij} \quad \text{discrete perimeter}$$

$$\delta(S) = \sum_{i \in S} \sum_{j \neq i} w_{ij} \quad \text{discrete perimeter}$$

Normalized cut problem

$$\min_{S \subset V} \frac{\sigma(S)}{\min\{\delta(S), \delta(S^c)\}} \quad \stackrel{\text{DEF}}{=} \quad h(G) \quad \text{Cheeger constant}$$

Normalized cut and Cheeger constant

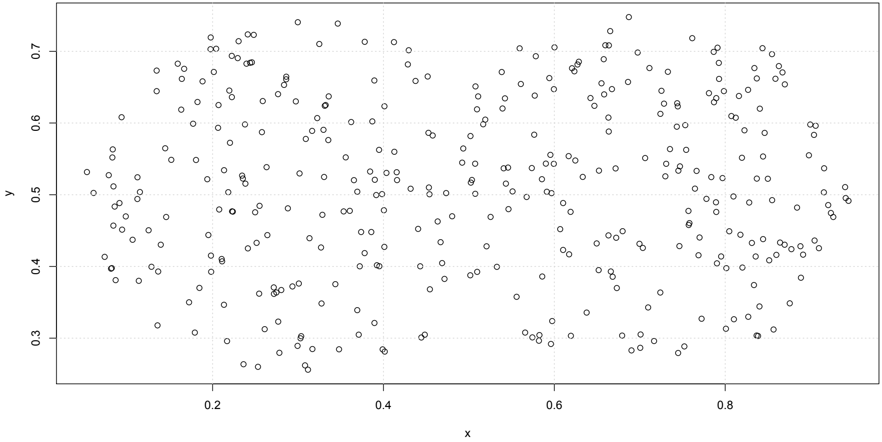
- The Cheeger constant is also called *conductance*.
- Small Cheeger constant $\equiv$ strong bottleneck.
- Best split set S defines a partition of the graph G.

Normalized cut and Cheeger constant

- The Cheeger constant is also called *conductance*.
- Small Cheeger constant $\equiv$ strong bottleneck.
- Best split set S defines a partition of the graph G.
- But the optimization problem NP-hard.

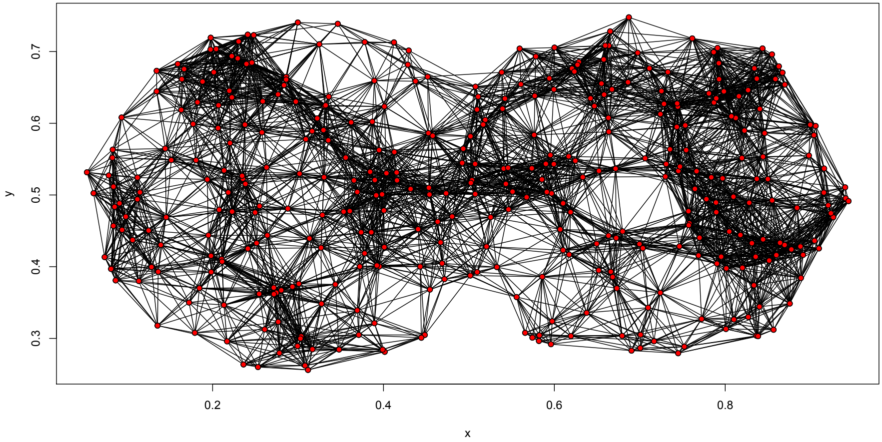
Example: observations

The set M is the union of two discs. ($n = 300$ points uniform from M.)



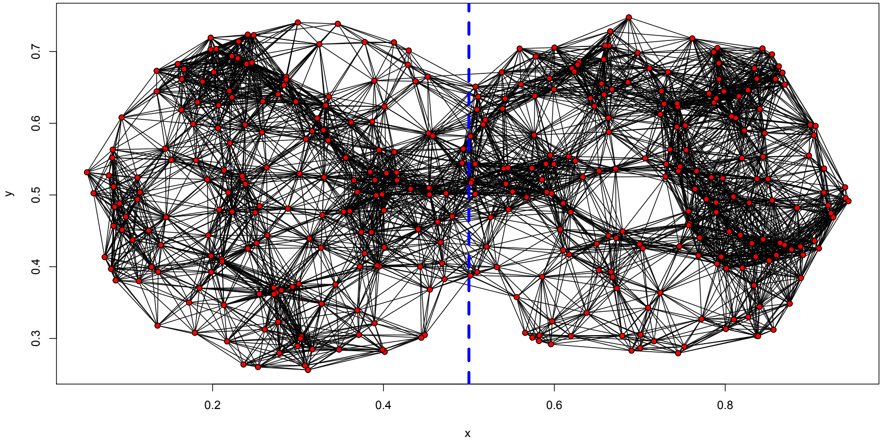
Example: graph

This is a neighborhood graph on the sample points.



Example: where to cut?

Finding a split that optimizes the normalized cut criterion is NP-hard.



Graph Laplacians and spectral graph partitioning

- Define the *degree matrix*

$$\mathbf{D} = \mathrm{diag}(\sum_j w_{ij},\ 1 \leq i \leq n).$$

- Define the *normalized graph Laplacian*

$$\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1}\mathbf{W}.$$

Graph Laplacians and spectral graph partitioning

- Define the *degree matrix*

$$\mathbf{D} = \mathrm{diag}(\sum_j w_{ij},\ 1 \leq i \leq n).$$

- Define the *normalized graph Laplacian*

$$\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1}\mathbf{W}.$$

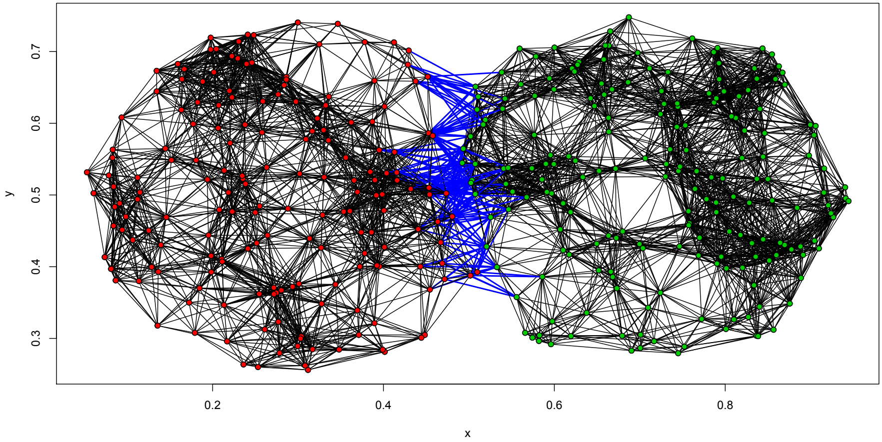
Graph bisection (e.g., Shi and Malik, 2000)

1. Compute the eigenvector for the second smallest eigenvalue of $\mathbf{L}$.
2. Partition the points according to their corresponding entry in this vector.

See also (Chung, 1997) and (Ng, Jordan, and Weiss, 2002).

Example: approximate best split

Partition computed using spectral bisection. (Blue: discrete boundary.)



The problem

Assuming the points $X_1, \ldots, X_n$ are sampled iid uniform from a domain $M \subset \mathbb{R}^d$, describe the large-sample behavior of the Cheeger constant of a ϵ_n-ball neighborhood graph.

- EAC, B. Pelletier, and P. Pudlo. The normalized graph cut and Cheeger constant: from discrete to continuous. *Adv. in Applied Probability*, 2012.

Literature

Closely related work:

- H. Narayanan, M. Belkin, and P. Niyogi. On the relation between low density separation, spectral clustering and graph cuts. *NIPS*, 2007.
- H. Narayanan and P. Niyogi. On the sample complexity of learning smooth cuts on a manifold. *COLT*, 2009.
- M. Maier, U. Von Luxburg, and M. Hein. Influence of graph construction on graph-based clustering measures. *NIPS*, 2009.
- M. Maier, U. von Luxburg, M. Hein. How the result of graph clustering methods depends on the construction of the graph. ESAIM: Probability and Statistics, 2013.

Setting

- $M \subset \mathbb{R}^d$ bounded, open and connected, with smooth boundary. (Assume that $\mathrm{Vol}_d(M) = 1$ without loss of generality.)
- $X_1, \ldots, X_n$ sampled iid uniformly from M.

Setting

- $M \subset \mathbb{R}^d$ bounded, open and connected, with smooth boundary. (Assume that $\mathrm{Vol}_d(M) = 1$ without loss of generality.)
- $X_1, \ldots, X_n$ sampled iid uniformly from M.

Smooth here means with positive reach. The reach of a set $A \subset \mathbb{R}^d$ is the supremum of all $r > 0$ such that, for all $x \in A \oplus B(0, r)$ there is a unique point $a \in \bar{A}$ such that

$$\|x - a\| = \min_{b \in A} \|x - b\|$$

See (Federer, 1959). (Related to the condition number of Niyogi et al.)

We consider the r_n-neighborhood graph $G_n = (V_n, E_n)$:

(i) vertices: $V_n = \{1, \ldots, n\}$

(ii) edges: $i \sim j$ if $\|X_i - X_j\| \leq r_n$

We consider the r_n-neighborhood graph $G_n = (V_n, E_n)$:

(i) vertices: $V_n = \{1, \dots, n\}$

(ii) edges: $i \sim j$ if $\|X_i - X_j\| \leq r_n$

Recall the Cheeger constant of the graph G_n:

$$h(G_n) = \min_{S \subset V_n} \frac{\sigma(S)}{\min\{\delta(S), \delta(S^c)\}}, \quad \text{with}$$

$$\sigma(S) = \sum_{i \in S} \sum_{j \in S^c} w_{ij} \quad \text{and} \quad \delta(S) = \sum_{i \in S} \sum_{j \neq i} w_{ij}$$

$$w_{ij} = \mathbf{1}_{\{\|X_i - X_j\| \leq r_n\}}$$

The continuous Cheeger constant

For $A \subset M$, set

$$\mu(A) = \mathrm{Vol}_d(A \cap M), \quad \nu(A) = \mathrm{Vol}_{d-1}(\partial A \cap M)$$

and define

$$h(A; M) = \frac{\nu(A)}{\min\{\mu(A), \mu(A^c)\}},$$

with Vol_k the k-dimensional Hausdorff measure.

The continuous Cheeger constant

For $A \subset M$, set

$$\mu(A) = \mathrm{Vol}_d(A \cap M), \quad \nu(A) = \mathrm{Vol}_{d-1}(\partial A \cap M)$$

and define

$$h(A; M) = \frac{\nu(A)}{\min\{\mu(A), \mu(A^c)\}},$$

with Vol_k the k-dimensional Hausdorff measure.

The Cheeger constant of M

$$h(M) = \inf\{h(A; M) \ : \ A \subset M\}.$$

The continuous Cheeger constant

For $A \subset M$, set

$$\mu(A) = \mathrm{Vol}_d(A \cap M), \quad \nu(A) = \mathrm{Vol}_{d-1}(\partial A \cap M)$$

and define

$$h(A; M) = \frac{\nu(A)}{\min\{\mu(A), \mu(A^c)\}},$$

with Vol_k the k-dimensional Hausdorff measure.

The Cheeger constant of M

$$h(M) = \inf\{h(A; M) \,:\, A \subset M\}.$$

- The minimization can be restricted to subsets A with smooth boundary of codimension 1.
- A Cheeger set $A^\star$ is a subset with $h(A^\star; M) = h(M)$.
- $\partial A^\star$ is not necessarily smooth (e.g., $d \geq 8$).

Relating the discrete and the continuous

A natural question...

As the sample size increases ($n \to \infty$) how is the (discrete) Cheeger constant $h(G_n)$ related to the (continuous) Cheeger constant $h(M)$?

Discrete perimeter and volume of a continuous set

For $A \subset \mathbb{R}^d$, let $S_A = \{i : X_i \in A\}$, and define

- the (normalized) discrete perimeter

$$\nu_n(A) = \frac{1}{\gamma_d r_n^{d+1}} \frac{1}{n(n-1)} \sigma_n(S_A)$$

where

$$\gamma_d = \int_{\mathbb{R}^d} \max\left(\langle u, z\rangle, 0\right) \mathbf{1}_{\{\|z\| \le 1\}} \, \mathrm{d}z,$$

where u is any unit-norm vector of $\mathbb{R}^d$.

Discrete perimeter and volume of a continuous set

For $A \subset \mathbb{R}^d$, let $S_A = \{i : X_i \in A\}$, and define

- the (normalized) discrete perimeter

$$\nu_n(A) = \frac{1}{\gamma_d r_n^{d+1}} \frac{1}{n(n-1)} \sigma_n(S_A)$$

  where

$$\gamma_d = \int_{\mathbb{R}^d} \max\left(\langle u, z\rangle, 0\right) \mathbf{1}_{\{\|z\| \leq 1\}} \, \mathrm{d}z,$$

  where u is any unit-norm vector of $\mathbb{R}^d$.

- the (normalized) discrete volume

$$\mu_n(A) = \frac{1}{\omega_d r_n^d} \frac{1}{n(n-1)} \delta_n(S_A)$$

  where ω_d denote the d-volume of the unit d-dimensional ball.

Discrete normalized cut of a continuous set

Define

$$h_n(A; G_n) = \frac{\nu_n(A)}{\min\{\mu_n(A), \mu_n(A^c)\}}$$

Discrete normalized cut of a continuous set

Define

$$h_n(A; G_n) = \frac{\nu_n(A)}{\min\{\mu_n(A), \mu_n(A^c)\}}$$

Theorem

Let $A \subset \mathbb{R}^d$ is such that $\partial A \cap M$ has positive reach. If $r_n \to 0$ with $nr_n^{d+1} / \log n \to \infty$, then

$$h_n(A; G_n) \to h(A; M) \quad a.s.$$

One side of the asymptotics

Corollary

If $r_n \to 0$ with $nr_n^{d+1} / \log n \to \infty$, then

$$\limsup_{n\to\infty} \frac{\omega_d}{\gamma_d} \frac{1}{r_n} h(G_n) \leq h(M) \quad a.s.$$

One side of the asymptotics

Corollary

If $r_n \to 0$ with $nr_n^{d+1}/\log n \to \infty$, then

$$\limsup_{n\to\infty} \frac{\omega_d}{\gamma_d}\frac{1}{r_n} h(G_n) \le h(M) \quad a.s.$$

Proof. This follows immediately from applying the previous result. Take $A \subset \mathbb{R}^d$ is such that $\partial A \cap M$ has positive reach. Then

$$h(G_n) \le \frac{\omega_d}{\gamma_d}\frac{1}{r_n} h_n(A; G_n) \to h(A; M)$$

This implies that

$$\limsup_n h(G_n) \le h(A; M)$$

for all such A. And minimizing the RHS over such A gives $h(M)$.

Concentration inequality for the discrete volume

Proposition

Fix a sequence $r_n \to 0$. Let $A \subset M$ be an arbitrary open subset of M. There exists a constant C depending only on M such that, for any $\varepsilon > 0$, and all n large enough, we have

$$\mathbb{P}\left[|\mu_n(A) - \mu(A)| \geq \varepsilon\right] \leq 2\exp\left(-\frac{nr_n^d \varepsilon^2}{C(1+\varepsilon)}\right).$$

In particular, if $nr_n^d / \log n \to \infty$, then $\mu_n(A) \to \mu(A)$ a.s. when $n \to \infty$.

By the triangle inequality, we have

$$\begin{aligned}|\mu_n(A) - \mu(A)| &\leq |\mu_n(A) - \mathbb{E}[\mu_n(A)]| + |\mathbb{E}[\mu_n(A)] - \mu(A)| \\ &= (1) + (2)\end{aligned}$$

By the triangle inequality, we have

$$\begin{aligned}|\mu_n(A) - \mu(A)| &\leq |\mu_n(A) - \mathbb{E}\left[\mu_n(A)\right]| + |\mathbb{E}\left[\mu_n(A)\right] - \mu(A)| \\ &= (1) + (2)\end{aligned}$$

Define the kernel

$$\phi_{A,r}(x, y) = \frac{1}{2}\Big\{\mathbf{1}_A(x) + \mathbf{1}_A(y)\Big\}\mathbf{1}\{\|x - y\| \leq r\}$$

so that $\mu_n(A)$ may be expressed as the following U-statistic

$$\mu_n(A) = \frac{1}{\omega_d n(n-1) r_n^d} \sum_{i \neq j} \phi_{A,r_n}(X_i, X_j)$$

By the triangle inequality, we have

$$\begin{aligned}|\mu_n(A) - \mu(A)| &\leq |\mu_n(A) - \mathbb{E}[\mu_n(A)]| + |\mathbb{E}[\mu_n(A)] - \mu(A)| \\ &= (1) + (2)\end{aligned}$$

Define the kernel

$$\phi_{A,r}(x, y) = \frac{1}{2}\Big\{\mathbf{1}_A(x) + \mathbf{1}_A(y)\Big\}\mathbf{1}\{\|x - y\| \leq r\}$$

so that $\mu_n(A)$ may be expressed as the following U-statistic

$$\mu_n(A) = \frac{1}{\omega_d n(n-1) r_n^d} \sum_{i \neq j} \phi_{A,r_n}(X_i, X_j)$$

We control (1) using a concentration inequality for U-statistics.

Define

$$M_r = \{x \in M : \mathrm{dist}(x, \partial M) > r\}$$

Define

$$M_r = \{x \in M : \text{dist}(x, \partial M) > r\}$$

Lemma

For any $A \subset M$ and $r < \mathrm{reach}(\partial M)$,

$$\left| \frac{1}{\omega_d r^d} \mathbb{E}\left[\phi_{A,r}(X_1, X_2)\right] - \mu(A) \right| \leq \mu(A \cap M_r^c)$$

Note that $\mathbb{E}[\mu_n(A)] = \frac{1}{\omega_d r_n^d} \mathbb{E}\left[\phi_{A,r_n}(X_1, X_2)\right]$.

Define

$$M_r = \{x \in M : \mathrm{dist}(x, \partial M) > r\}$$

Lemma

For any $A \subset M$ and $r < \mathrm{reach}(\partial M)$,

$$\left| \frac{1}{\omega_d r^d} \mathbb{E}\left[\phi_{A,r}(X_1, X_2)\right] - \mu(A) \right| \leq \mu(A \cap M_r^c)$$

Note that $\mathbb{E}[\mu_n(A)] = \frac{1}{\omega_d r_n^d} \mathbb{E}\left[\phi_{A,r_n}(X_1, X_2)\right]$.

Proof. We have

$$\mathbb{E}\left[\phi_{A,r}(X_1, X_2)\right] = \mathbb{E}\left[\mathbf{1}_A(X_1)\mathbf{1}_{\{\|X_1 - X_2\| \leq r\}}\right].$$

Define

$$M_r = \{x \in M : \text{dist}(x, \partial M) > r\}$$

Lemma

For any $A \subset M$ and $r < \text{reach}(\partial M)$,

$$\left| \frac{1}{\omega_d r^d} \mathbb{E}\left[\phi_{A,r}(X_1, X_2)\right] - \mu(A) \right| \leq \mu(A \cap M_r^c)$$

Note that $\mathbb{E}[\mu_n(A)] = \frac{1}{\omega_d r_n^d} \mathbb{E}\left[\phi_{A,r_n}(X_1, X_2)\right]$.

Proof. We have

$$\mathbb{E}\left[\phi_{A,r}(X_1, X_2)\right] = \mathbb{E}\left[\mathbf{1}_A(X_1)\mathbf{1}_{\{\|X_1 - X_2\| \leq r\}}\right].$$

Conditioning on X_1, we have

$$\begin{aligned} \mathbb{E}\left[\mathbf{1}_{A\cap M_r}(X_1)\mathbf{1}_{\{\|X_1 - X_2\| \leq r\}}\right] &= \omega_d r^d \mu(A \cap M_r) \\ &= \omega_d r^d \mu(A) - \omega_d r^d \mu(A \cap M_r^c) \\ \mathbb{E}\left[\mathbf{1}_{A\cap M_r^c}(X_1)\mathbf{1}_{\{\|X_1 - X_2\| \leq r\}}\right] &\leq \omega_d r^d \mu(A \cap M_r^c). \end{aligned}$$

We control (2) — the bias — using this lemma and the following result, closely related to Weyl's volume formula for tubular neighborhoods.

Lemma

For any bounded open subset $R \subset \mathbb{R}^d$ with $\operatorname{reach}(\partial R) = \rho > 0$ and any $0 < r < \rho$,

$$\mathrm{Vol}_d(\mathcal{V}(\partial R, r)) \leq 2^d \, \mathrm{Vol}_{d-1}(\partial R) \, r.$$

This implies that

$$\mu(A \cap M_r^c) \leq \mu(M_r^c) \leq \mathrm{Vol}_d(\partial M, r) \leq Cr$$

for a constant $C = C(M)$.

Concentration inequality for the discrete perimeter

Proposition

Fix a sequence $r_n \to 0$. Let A be an open subset of M such that $\partial A \cap M$ has positive reach. There exists a constant C depending only on M such that, for any $\varepsilon > 0$, and for all n large enough, we have

$$\mathbb{P}\left[|\nu_n(A) - \nu(A)| \geq \epsilon\right] \leq 2\exp\left(-\frac{nr_n^{d+1}\epsilon^2}{C(\nu(A)+\epsilon)}\right).$$

In particular, if $nr_n^{d+1}/\log n \to \infty$, then $\nu_n(A) \to \nu(A)$ a.s. when $n \to \infty$.

Proof. The proof is analogous to that of the previous proposition (for the volume). Indeed, we can express $\nu_n(A)$ as a U-statistic

$$\nu_n(A) = \frac{1}{\gamma_d n(n-1) r_n^{d+1}} \sum_{i \neq j} \bar{\phi}_{A,r_n}(X_i, X_j),$$

where

$$\bar{\phi}_{A,r}(x,y) = \frac{1}{2}\Big\{\mathbf{1}_A(x)\mathbf{1}_{A^c}(y) + \mathbf{1}_A(y)\mathbf{1}_{A^c}(x)\Big\}\mathbf{1}\{\|x-y\| \leq r\}$$

The control of the bias is more delicate. We use the following bound.

Lemma

Let $A = R \cap M$, where R is a bounded domain with $\mathrm{reach}(\partial R) = \rho > 0$. Let $r < \min\{\rho/2, \mathrm{reach}(\partial M)\}$. There exists a constant $C = C(M) > 0$ such that

$$\begin{aligned}\left|\frac{1}{\gamma_d r^{d+1}}\mathbb{E}\left[\bar{\phi}_{A,r}(X_1,X_2)\right] - \nu(A)\right| &\leq C\,\mathrm{Vol}_{d-1}(\partial R \cap (\partial M \oplus B(0,r))) \\ &\quad + C\,\mathrm{Vol}_{d-1}(\partial R \cap M)\frac{r}{\rho}\end{aligned}$$

Note that

$$\mathbb{E}\left[\nu_n(A)\right] = \frac{1}{\gamma_d r_n^{d+1}}\mathbb{E}\left[\bar{\phi}_{A,r_n}(X_1,X_2)\right]$$

The control of the bias is more delicate. We use the following bound.

Lemma

Let $A = R \cap M$, where R is a bounded domain with $\mathrm{reach}(\partial R) = \rho > 0$. Let $r < \min\{\rho/2, \mathrm{reach}(\partial M)\}$. There exists a constant $C = C(M) > 0$ such that

$$\left|\frac{1}{\gamma_d r^{d+1}} \mathbb{E}\left[\bar{\phi}_{A,r}(X_1, X_2)\right] - \nu(A)\right| \leq C\,\mathrm{Vol}_{d-1}(\partial R \cap (\partial M \oplus B(0,r))) + C\,\mathrm{Vol}_{d-1}(\partial R \cap M)\frac{r}{\rho}$$

Note that

$$\mathbb{E}\left[\nu_n(A)\right] = \frac{1}{\gamma_d r_n^{d+1}} \mathbb{E}\left[\bar{\phi}_{A,r_n}(X_1, X_2)\right]$$

Applying the lemma, for $A = R \cap M$, we have

$$|\mathbb{E}\left[\nu_n(A)\right] - \nu(A)| \leq C\,\mathrm{Vol}_{d-1}(\partial R \cap (\partial M \oplus B(0,r_n))) + C\,\mathrm{Vol}_{d-1}(\partial A \cap M)\frac{r_n}{\mathrm{reach}(\partial R)} \to 0$$

Does the discrete converge to the continuous?

Do we have the counterpart to the corollary, meaning

Is it true that, for some $r_n \to 0$, we have

$$\frac{\omega_d}{\gamma_d} \frac{1}{r_n} h(G_n) \to h(M) \quad \text{a.s.} \quad n \to \infty?$$

Does the discrete converge to the continuous?

Do we have the counterpart to the corollary, meaning

Is it true that, for some $r_n \to 0$, we have

$$\frac{\omega_d}{\gamma_d} \frac{1}{r_n} h(G_n) \to h(M) \quad \text{a.s.} \quad n \to \infty?$$

Look at the following recent work:

- N. García Trillos and D. Slepcev. Γ-Convergence of Perimeter on Random Geometric Graphs. *CMU preprint*, 2013.
- N. García Trillos and D. Slepcev. Continuum limit of total variation on point clouds. *arXiv preprint*, 2014.

Consistent estimation

- The class of all open subsets of M with positive reach s too rich for us to obtain uniform convergences for the discrete volume and perimeter.

Consistent estimation

- The class of all open subsets of M with positive reach s too rich for us to obtain uniform convergences for the discrete volume and perimeter.
- Without loss of generality, assume that $M \subset [0, 1]^d$. We consider the class $\mathcal{R}_n$ of open subsets R of $[0, 1]^d$ with $\mathrm{reach}(\partial R) \geq \rho_n$ for a sequence $\rho_n \to 0$.

Consistent estimation of the Cheeger constant

Theorem

If

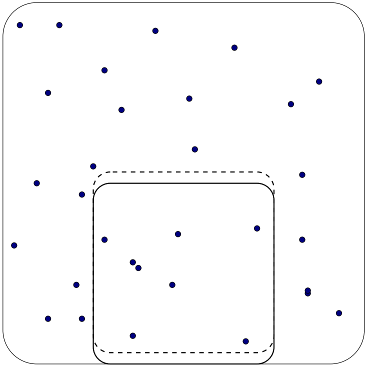
(i) $r_n \to 0$ *and* $nr_n^{2d+1} \to \infty$, *and*

(ii) $\rho_n \to 0$ *slowly with* $r_n = o(\rho_n^\alpha)$ *and* $nr_n^{2d+1}\rho_n^\alpha \to \infty$ *for all* $\alpha > 0$,

then

$$\min_{R \in \mathcal{R}_n} h_n(R; G_n) \to h(M) \quad a.s. \quad n \to \infty.$$

- The ingredients are uniform versions of the concentration inequalities for the discrete volume and perimeter over the class $\mathcal{R}_n$, obtained via the union bound and a bound on the covering number of $\mathcal{R}_n$.
- However, the bias for the discrete perimeter cannot be controlled uniformly over sets in $\mathcal{R}_n$.



Our way around that is to compare the discrete perimeter $\nu_n(R)$ with $\mathrm{Vol}_{d-1}(\partial R \cap M_{r_n})$ instead.
We get the following.

Lemma

Under the conditions of last theorem, we have

$$\liminf_{n\to\infty} \inf_{R\in\mathcal{R}_n} (h_n(R) - h(R; M_{r_n})) \geq 0 \quad a.s.$$

Proof of the last theorem.
For each n, take $R_n \in \mathcal{R}_n$. Then

$$\begin{aligned} h_n(R_n; G_n) - h(M) &= [h_n(R_n; G_n) - h(R_n; M_{r_n})] \\ &\quad + [h(R_n; M_{r_n}) - h(M_{r_n})] + [h(M_{r_n}) - h(M)] \\ &\geq \inf_{R \in \mathcal{R}_n} (h_n(R; G_n) - h(R; M_{r_n})) + [h(M_{r_n}) - h(M)] \end{aligned}$$

Proof of the last theorem.
For each n, take $R_n \in \mathcal{R}_n$. Then

$$\begin{aligned} h_n(R_n; G_n) - h(M) &= [h_n(R_n; G_n) - h(R_n; M_{r_n})] \\ &\quad + [h(R_n; M_{r_n}) - h(M_{r_n})] + [h(M_{r_n}) - h(M)] \\ &\geq \inf_{R \in \mathcal{R}_n} (h_n(R; G_n) - h(R; M_{r_n})) + [h(M_{r_n}) - h(M)] \end{aligned}$$

We have the following continuity property of the Cheeger constant.

Lemma

Under our conditions on M, $h(M_r) = (1 + O(r))h(M)$ as $r \to 0$.

We conclude that $\liminf_n \min_{R \in \mathcal{R}_n} h_n(R_n; G_n) \geq h(M)$.

Proof of the last theorem.

For each n, take $R_n \in \mathcal{R}_n$. Then

$$\begin{aligned} h_n(R_n; G_n) - h(M) &= [h_n(R_n; G_n) - h(R_n; M_{r_n})] \\ &\quad + [h(R_n; M_{r_n}) - h(M_{r_n})] + [h(M_{r_n}) - h(M)] \\ &\geq \inf_{R \in \mathcal{R}_n} (h_n(R; G_n) - h(R; M_{r_n})) + [h(M_{r_n}) - h(M)] \end{aligned}$$

We have the following continuity property of the Cheeger constant.

Lemma

Under our conditions on M, $h(M_r) = (1 + O(r))h(M)$ as $r \to 0$.

We conclude that $\liminf_n \min_{R \in \mathcal{R}_n} h_n(R_n; G_n) \geq h(M)$.

For an upper bound, use the first theorem.

Consistent estimation of Cheeger sets

Theorem

Let $R_n \in \operatorname{argmin}_{R \in \mathcal{R}_n} h_n(R; G_n)$. Then, with probability one:

(*i*) $\{R_n \cap M\}$ *admits a subsequence converging in* L^1;

(*ii*) *any convergent subsequence of* $\{R_n \cap M\}$ *converges to a Cheeger set in* L^1.

Consistent estimation of Cheeger sets

Theorem

Let $R_n \in \mathrm{argmin}_{R \in \mathcal{R}_n} h_n(R; G_n)$. Then, with probability one:

(*i*) $\{R_n \cap M\}$ *admits a subsequence converging in* L^1;

(*ii*) *any convergent subsequence of* $\{R_n \cap M\}$ *converges to a Cheeger set in* L^1.

The problem here is that we do not know M, so that $R_n \cap M$ is not a valid estimator. (More on that later.)

L^1 metric on Borel sets

- For A and B Borel subsets of $\mathbb{R}^d$:

$$\int |\mathbf{1}_A(x) - \mathbf{1}_B(x)| \, \mathrm{d}x = \mathrm{Vol}_d \left(A \Delta B \right).$$

L^1 metric on Borel sets

- For A and B Borel subsets of $\mathbb{R}^d$:

$$\int |\mathbf{1}_A(x) - \mathbf{1}_B(x)| \, \mathrm{d}x = \mathrm{Vol}_d \left(A \Delta B \right).$$

- de Giorgi perimeter of Ω, measurable subset of M:

$$P_M(\Omega) = \sup \left\{ \int_\Omega \mathrm{div}(\varphi) \mathrm{d}x \ : \ \varphi \in \mathcal{C}_c^\infty(M; \mathbb{R}^d), \ \|\varphi\|_\infty \leq 1 \right\}.$$

L^1 metric on Borel sets

- For A and B Borel subsets of $\mathbb{R}^d$:

$$\int |\mathbf{1}_A(x) - \mathbf{1}_B(x)| \,\mathrm{d}x = \mathrm{Vol}_d\,(A \Delta B)\,.$$

- de Giorgi perimeter of Ω, measurable subset of M:

$$P_M(\Omega) = \sup \left\{ \int_\Omega \mathrm{div}(\varphi)\mathrm{d}x \;:\; \varphi \in \mathcal{C}_c^\infty(M; \mathbb{R}^d),\ \|\varphi\|_\infty \leq 1 \right\}.$$

- $P_M(\Omega) = \mathrm{Vol}_{d-1}(\partial\Omega \cap M)$ for Ω of class $\mathcal{C}^1$.

L^1 metric on Borel sets

Proposition (Compactness)

Let (E_n) be a sequence of measurable subsets of M such that

$$\limsup_{n\to\infty} P_M(E_n) < \infty.$$

Then (E_n) admits a subsequence converging for the L^1 metric.

L^1 metric on Borel sets

Proposition (Compactness)

Let (E_n) be a sequence of measurable subsets of M such that

$$\limsup_{n\to\infty} P_M(E_n) < \infty.$$

Then (E_n) admits a subsequence converging for the L^1 metric.

Proposition (Lower semi-continuity)

Let (E_n) and E be measurable subsets of M such that $E_n \xrightarrow{L^1} E$. Then

$$\lim_{n\to\infty} \mathrm{Vol}_d(E_n) \to \mathrm{Vol}_d(E) \quad \textit{and} \quad \liminf_{n\to\infty} P_M(E_n) \geq P_M(E).$$

See (Giusti, 1984) or (Henrot and Pierre, 2005).

Define the probability measure

$$Q_n = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{R_n}(X_i)\delta_{X_i}$$

Note that Q_n can be computed from the data.

Define the probability measure

$$Q_n = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{R_n}(X_i) \delta_{X_i}$$

Note that Q_n can be computed from the data.

Theorem

Almost surely, any accumulation point of $\{Q_n\}$ is of the form $Q = \mathbf{1}_{A_\infty} \mu$ with A_∞ a Cheeger set of M.

Define the probability measure

$$Q_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{R_n}(X_i)\delta_{X_i}$$

Note that Q_n can be computed from the data.

Theorem

Almost surely, any accumulation point of $\{Q_n\}$ is of the form $Q = \mathbf{1}_{A_\infty}\mu$ with A_∞ a Cheeger set of M.

It is possible to reconstruct a Cheeger set of M from the discrete measure Q_n. It amounts to estimating its support. For example, one can take a union of small balls around each point in R_n.

Numerical approximation: spectral clustering

- Computing a normalized cut is NP-hard. Our method is not computationally tractable.
- Is spectral clustering consistent?